Έργο:	**«ΘΑΛΗΣ: Ενίσχυση της Διεπιστημονικής ή και Διιδρυματικής έρευνας και καινοτομίας με δυνατότητα προσέλκυσης ερευνητών υψηλού επιπέδου από το εξωτερικό μέσω της διενέργειας βασικής και εφαρμοσμένης έρευνας αριστείας»**
Τίτλος Υποέργου:	**«ΕΙΚΟΣ»: Θεωρητική και αλγοριθμική θεμελίωση για Προσωποκεντρικά Συνεργατικά Πληροφοριακά Συστήματα**

Παραδοτέο Π.4.1

Διαχείριση πληροφορίας σε μεγάλης κλίμακας πληροφορικά συστήματα

Σεπτέμβριος 2015

Ευρωπαϊκή Ένωση
Ευρωπαϊκό Κοινωνικό Ταμείο
ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΕΚΠΑΙΔΕΥΣΗ ΚΑΙ ΔΙΑ ΒΙΟΥ ΜΑΘΗΣΗ
επένδυση στην κοινωνία της γνώσης
ΥΠΟΥΡΓΕΙΟ ΠΑΙΔΕΙΑΣ & ΘΡΗΣΚΕΥΜΑΤΩΝ, ΠΟΛΙΤΙΣΜΟΥ & ΑΘΛΗΤΙΣΜΟΥ
ΕΙΔΙΚΗ ΥΠΗΡΕΣΙΑ ΔΙΑΧΕΙΡΙΣΗΣ
ΕΣΠΑ 2007-2013
ΕΥΡΩΠΑΪΚΟ ΚΟΙΝΩΝΙΚΟ ΤΑΜΕΙΟ
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Δράση 4	**Κατανεμημένες Υποδομές Αποθήκευσης, Προσπέλασης και Διαχείρισης Δεδομένων**				
Ομάδα	Ερ. Ομάδα 4	**Έναρξη**	01/06/2012	**Λήξη**	30/11/2015
Συντονιστής ΕΟ4	Παναγιώτης Τριανταφύλλου (Παν. Πατρών)				
Υποδράση: ΥΔ 4.1	Διαχείριση πληροφορίας σε μεγάλης κλίμακας πληροφορικά συστήματα				
Συμμετέχοντες	*Μέλη ΚΕΟ*	Παναγιώτης Τριανταφύλλου (Παν. Πατρών), Αντώνιος Δεληγιαννάκης (Πολυτεχνείο Κρήτης) , Βασίλειος Σαμολαδάς (Πολυτεχνείο Κρήτης), Ιωάννης Κωτίδης (ΟΠΑ)			
	Μέλη ΟΕΣ	Δημήτρης Καραμπίνας (Παν. Πατρών), Δημήτρης Μπούσης (Παν. Πατρών), Κυριακή Παναγίδη (Παν. Πατρών), Δημήτρης Σαχαρίδης (ΙΠΣΥ – Ε.Κ. ΑΘΗΝΑ), Κωνσταντίνα Μακρυνιώτη (ΟΠΑ), Κωνσταντίνος Γεωργούλας (ΟΠΑ), Κωνσταντίνος Ζαγγανάς (ΙΠΣΥ – Ε.Κ. ΑΘΗΝΑ), Γεώργιος Ζώης (ΟΠΑ)			

Σύντομη Περιγραφή	Η Υποδράση ΥΔ 4.1 επικεντρώνεται στη διαχείριση πληροφορίας σε μεγάλης κλίμακας πληροφορικά συστήματα. Οι κύριοι άξονες της υποδράσης αφορούν (ι) στη σχεδίαση αποδοτικών τεχνικών εξατομικευμένης επεξεργασίας σύνθετων επερωτήσεων σε μεγάλης κλίμακας εφαρμογές, (ιι) στην αξιοποίηση ανθρωποκεντρικών τεχνικών δεικτοδότησης για την απρόσκοπτη πρόσβαση στα δεδομένα τέτοιων συστημάτων και (ιιι) στην αντιμετώπιση περιπτώσεων ελλιπούς πληροφορίας μέσω της αντικατάστασης αγνοούμενων τιμών. Οι προτεινόμενες τεχνικές και αλγόριθμοι αξιοποιούν σύγχρονα συστήματα παράλληλης και κατανεμημένης επεξεργασίας, επιτρέποντας την αξιοποίηση τους σε μεγάλης κλίμακας πληροφοριακά συστήματα.
Παραδοτέο	Π.4.1 Διαχείριση πληροφορίας σε μεγάλης κλίμακας πληροφορικά συστήματα
/Στόχος στο Τ.Δ./	Τεχνική αναφορά που θα περιλαμβάνει τουλάχιστον 2 δημοσιεύσεις.
Επίτευξη στόχου	100%

Περιεχόμενα

1 Εισαγωγή

Κεντρικός στόχος του έργου ΕΙΚΟΣ είναι να προσφέρει τη μεθοδολογία, τη θεωρητική θεμελίωση, τις αλγοριθμικές τεχνικές και την αρχιτεκτονική του λογισμικού που απαιτείται ώστε τα πληροφοριακά συστήματα να μπορούν να προσφέρουν στους χρήστες αφενός την δυνατότητα εξατομίκευσης της παρεχόμενης πληροφορίας και αφετέρου τη δυνατότητα χρήσης ενσωματωμένων ετερογενών δεδομένων, ενδεχομένως διαφορετικής προέλευσης, με διαφανή τρόπο.

Στα πλαίσια του έργου, η Δράση 4 με τίτλο «Κατανεμημένες Υποδομές Αποθήκευσης, Προσπέλασης και Διαχείρισης Δεδομένων» σκοπό έχει να παράσχει αρχιτεκτονικές και αλγορίθμους οι οποίοι είτε οι ίδιες θα παρέχουν τρόπους για την κατανεμημένη οργάνωση χώρων δεδομένων στα χαμηλότερα επίπεδα του συστήματος, είτε θα παρέχουν κατάλληλα στατιστικά στοιχεία (όπως συνόψεις δεδομένων που περιγράφουν το φόρτο του συστήματος, ή το ποια δεδομένα ζητούνται από ποιούς χρήστες, τι ρόλο παίζουν διάφοροι χρήστες, κλπ) που να υποβοηθούν την οργάνωση αυτή. Επιπλέον, στόχος είναι οι ίδιοι οι χρήστες να αναβαθμιστούν από απλοί καταναλωτές πληροφορίας σε πρωταγωνιστές που μέσω της συνεργατικότητάς τους να βοηθούν στην κατανόηση των δεδομένων, της σχέσης τους και στην εξατομίκευση των αποτελεσμάτων με βάση το ποιός ερωτά και την «κοινοτική σοφία» αναφορικά με τα περιεχόμενα του οικοσυστήματος.

Η Δράση 4 οργανώνεται στις εξής υποδράσεις: ΥΔ 4.1 Διαχείριση πληροφορίας σε μεγάλης κλίμακας πληροφορικά συστήματα, ΥΔ 4.2 Μέθοδοι για παραγωγή στατιστικών και συνόψεων δεδομένων και ΥΔ 4.3 Μέθοδοι εύρεσης ομοιότητας χρηστών υπερχώρων δεδομένων.

Το παρόν Παραδοτέο Π.4.1 περιλαμβάνει τα αποτελέσματα της υποδράσης ΥΔ 4.1. Στην ενότητα **Error! Reference source not found.** παρουσιάζουμε το γενικότερο πλαίσιο του προβλήματος και αναφέρουμε τους κύριους άξονες γύρω από τους οποίους οργανώσαμε την έρευνα μας. Στην ενότητα 3

παρουσιάζουμε συνοπτικά τις τεχνικές και αλγορίθμους οι οποίες προέκυψαν για την υλοποίηση των στόχων. Τέλος, ανακεφαλαιώνουμε τα αποτελέσματά μας στην ενότητα 4.

2 Διαμόρφωση ερευνητικού πλαισίου

Η Δράση 4 «Κατανεμημένη Υποδομή για Αποθήκευση, Πρόσβαση και Διαχείριση Δεδομένων» απαντά στο ερώτημα: «Πώς αντιμετωπίζουμε την κλιμάκωση του συστήματος (σε αριθμό χρηστών, όγκο δεδομένων και μέγεθος κατανεμημένων υποδομών) μέσω κατανεμημένων τεχνικών;». Σκοπός είναι ο σχεδιασμός αρχιτεκτονικών, αλγορίθμων και υποβοηθητικών δομών δεδομένων για την κατανεμημένη οργάνωση και επεξεργασία/επερώτηση των δεδομένων του οικοσυστήματος.

Κεντρικό ζητούμενο στα πλαίσια της υποδράσης είναι να μελετηθούν τεχνικές διαχείρισης των δεδομένων, μέσω κατάλληλων μοντέλων επερωτήσεων οι οποίες θα επιτρέπουν τις παρεμβάσεις χρηστών στο τρόπο με τον οποίον τα συστήματα αξιολογούν και προβάλλουν τα δεδομένα, με τρόπο που να εξυπηρετεί τους στόχους τους. Με αυτό τον τρόπο παρέχουμε ανεξαρτησία από το μοντέλο περιγραφής των δεδομένων και επιτυγχάνουμε την προσωποποιημένη πρόσβαση σε αυτά μέσω κατάλληλων τεχνικών αναζήτησης και επεξεργασίας. Επιπλέον ζητούμενο είναι να εστιάσουμε σε σύγχρονες κατανεμημένες αρχιτεκτονικές και αλγόριθμους. Για την αξιοποίηση των στόχων της ΥΔ 4.1 ορίστηκαν 3 κύριοι άξονες γύρω από τους οποίους εστιάσαμε την έρευνα μας.

Ένας βασικός άξονας της ΥΔ 4.1 είναι διαχείριση πληροφορίας σε μεγάλης κλίμακας πληροφορικά συστήματα μέσω της σχεδίασης αποδοτικών αλγορίθμων εξατομικευμένης επεξεργασίας σύνθετων επερωτήσεων. Η κλιμάκωση του όγκου και της ποσότητας των δεδομένων στα σύγχρονα πληροφοριακά συστήματα κάνει επιτακτική την πρόσβαση σε αυτά μέσω τεχνικών που καθοδηγούν το χρήστη στα ποιο χρήσιμα δεδομένα, επιτρέποντας έτσι την αξιοποίηση του πλούτου της παρεχόμενης πληροφορίας χωρίς το κίνδυνο αυτός να χαθεί σε ένα πέλαγος υπέρ-πληροφόρησης.

Στα πλαίσια του στόχου αυτού μελετήθηκαν εκτενώς σύνθετα επερωτήματα τα οποία επιτρέπουν την ανάλυση προτιμήσεων χρηστών σε μεγάλης κλίμακας οικοσυστήματα και την αξιοποίηση των προτιμήσεων αυτών για την παροχή ταξινομημένων και εξατομικευμένων αποτελεσμάτων σε σύνθετες επερωτήσεις. Παραδείγματα τέτοιων επερωτήσεων είναι τα ερωτήματα κορυφογραμμής (skyline queries), τα ερωτήματα κατάταξης (top-k queries) και τα ερωτήματα ταξινόμησης-σύζευξης (rank-joins) καθώς και συνδυασμοί τους.

Ένας δεύτερος άξονας της ΥΔ 4.1 είναι η δημιουργία αποδοτικών τεχνικών δεικτοδότησης για την απρόσκοπτη διαχείριση δεδομένων που προέρχονται από ομάδες χρηστών κατανεμημένων σε διαφορετικές γεωγραφικά περιοχές. Βασικό ζητούμενο είναι η επεξεργασία των δεδομένων αυτών μέσω ανθρωποκεντρικών τεχνικών μέσα σε περιβάλλοντα συνεργασίας. Για το σκοπό αυτό δόθηκε έμφαση στη δημιουργία ανθρωποκεντρικών δομών δεικτοδότησης αξιοποιώντας ετικέτες (tags) επισήμανσης για τη δημιουργία ταξονομιών μέσω τεχνικών crowdsourcing.

Ο τρίτος ερευνητικός άξονας της ΥΔ 4.1 αφορά την αντιμετώπιση ενός επιτακτικού προβλήματος το οποίο παρουσιάζεται στις περισσότερες υλοποιήσεις μεγάλης κλίμακας πληροφορικών συστημάτων και έχει να κάνει με το πρόβλημα της αντιμετώπισης αγνοούμενων τιμών. Η έρευνα μας επικεντρώθηκε στην ανάπτυξη τεχνικών αντικατάστασης αγνοούμενων τιμών οι οποίες θα μπορούν ταυτόχρονα να αξιοποιήσουν τις αναδυόμενες τεχνολογίες συστημάτων μεγάλων δεδομένων.

3 Αναλυτικά Αποτελέσματα

3.1 Τεχνικές εξατομικευμένης επεξεργασίας σύνθετων επερωτήσεων

3.1.1 Ανίχνευση προτιμώμενων δεδομένων σε μεγάλης κλίμακας οικοσυστήματα

Οι επερωτήσεις κορυφογραμμής (skyline queries) αποτελούν βασικό εργαλείο για την ανάλυση δεδομένων προτιμήσεων χρηστών σε οικοσυστήματα και σε κοινωνικά δίκτυα. Στην εργασία [ArDV12] παρουσιάζουμε καινοτόμους αλγόριθμους για την αποδοτική επεξεργασία δύο σημαντικών κλάσεων επερωτήσεων που αφορούν προτιμήσεις χρηστών. Δεδομένου ενός συνόλου προτιμήσεων χρηστών και ενός συνόλου πιθανών υποψηφίων δεδομένων (πχ προϊόντων σε ένα οικοσύστημα), σκοπός του πρώτου προβλήματος αποτελεί η ταχεία εύρεση όλων των χρηστών που θεωρούν ένα συγκεκριμένο δεδομένο ή προϊόν ως το πιο προτιμητέο. Αντίστοιχα, σκοπός του δεύτερου προβλήματος είναι η ανίχνευση των k πιο προτιμητέων δεδομένων ή προϊόντων από όλους τους χρήστες.

Όσον αφορά το πρώτο πρόβλημα, ορίσαμε την ελκυστικότητα ενός δεδομένου ή προϊόντος βασισμένοι σε ανεστραμμένες επερωτήσεις κορυφογραμμής (reverse skyline queries). Μετά παρουσιάζουμε ένα νέο αλγόριθμο, τον RSA, ο οποίος μειώνει σημαντικά το κόστος εισόδου/εξόδου (I/O) ως προς τον προηγούμενο καλύτερο αλγόριθμο BRS, ενώ την ίδια στιγμή μπορεί γρήγορα να παρέχει τα πρώτα αποτελέσματα.

Για την επεξεργασία ενός μεγάλου αριθμού επερωτήσεων που απαιτεί το δεύτερο πρόβλημα που μελετάμε, αναπτύξαμε επίσης μία ομαδοποιημένη και κατανεμημένη επέκταση του RSA αλγόριθμου. Ο νέος μας αλγόριθμος παρέχει σημαντικά πιο γρήγορη επεξεργασία επερωτήσεων με την εκμετάλλευση κοινών απαιτήσεων για είσοδο/έξοδο και τη δυνατότητα κοινής επεξεργασίας των προτιμήσεων των χρηστών. Η πειραματική μας μελέτη χρησιμοποιώντας τόσο πραγματικά, όσο και συνθετικά δεδομένα αποδεικνύει την κυριαρχία των αλγόριθμών μας για τις υπό μελέτη κλάσεις επερωτήσεων.

Οι προτεινόμενες τεχνικές επιτρέπουν την απάντηση πολλαπλών επερωτήσεων κορυφογραμμής σε μεγάλης κλίμακας οικοσυστήματα, δεδομένων κάποιων προτιμήσεων χρηστών. Ο προτεινόμενοι αλγόριθμοι είναι πολύ πιο αποδοτικοί από προηγούμενες τεχνικές για μεγάλης κλίμακας δεδομένα και έχουν σχεδιαστεί ώστε να επιτρέπεται η εκτέλεση τους παράλληλα και κατανεμημένα, σύμφωνα με τις απαιτήσεις της ΥΔ 4.1.

Μια άλλη άμεση εφαρμογή των προτεινόμενων τεχνικών είναι η επιλογή, ανάμεσα σε όλα τα δεδομένα του οικοσυστήματος, των πιο προτιμητέων από τους χρήστες δεδομένων και η χρήση αυτής της πληροφορίας για την βελτίωση της απόκρισης σε μελλοντικές επερωτήσεις μέσω της αποθήκευσής τους (caching).

3.1.2 Επεξεργασία ερωτημάτων κατάταξης

Στα πλαίσια της υπό-δράσης ΥΔ 4.1 εξετάσαμε την επεξεργασία ερωτημάτων κατάταξης (rank queries) για μεγάλης κλίμακας κατανεμημένα οικοσυστήματα στα οποία οι κόμβοι εισέρχονται κι αποχωρούν δυναμικά. Τα ερωτήματα αυτά αναφέρονται σε μία μεγάλη κατηγορία τα οποία επιτάσσουν μία ολική διάταξη των πλειάδων και κατά συνέπεια την ανάκτηση των καλύτερων εκ των διατεταγμένων στοιχείων.

Η εργασία [TsSS14] προτείνει ένα γενικό πλαίσιο επεξεργασίας τέτοιων ερωτημάτων κατάλληλο για κατανεμημένα οικοσυστήματα τα οποία ακολουθούν τη δημοφιλή τεχνική του κατανεμημένου πίνακα κατακερματισμού (distributed hash table). Στα πλαίσια αυτά, η εξετάσαμε τρεις κατηγορίες ερωτημάτων.

Η πρώτη κατηγορία αφορά τα top-k ερωτήματα τα οποία επιβάλλουν μία διάταξη του πεδίου ορισμού (domain) μέσω μία μονότονης συνάρτησης. Το αποτέλεσμα περιλαμβάνει k πλειάδες οι οποίες αποτιμούνται υψηλότερα βάσει των κριτηρίων του χρήστη όταν συγκρίνονται με οποιαδήποτε άλλη αποθηκευμένη εγγραφή στο οικοσύστημα.

Η δεύτερη κατηγορία είναι τα ερωτήματα κορυφογραμμής τα οποία επιβάλλουν μία μερική διάταξη του πεδίου ως ορίζεται από τη συνάρτηση Pareto aggregation που προσδιορίζεται για κάθε χαρακτηριστικό ξεχωριστά (όσον αφορά τη μερική διάταξη, δύο εγγραφές μπορεί να μην δύναται να συγκριθούν μεταξύ τους). Η απάντηση σε ένα ερώτημα κορυφογραμμής είναι ένα σύνολο από στοιχεία που μεγιστοποιούν αυτή τη σχέση μερικής διάταξης. Μια σημαντική διαφορά ανάμεσα στις δύο αυτές κατηγορίες ερωτημάτων είναι ότι για τη διάταξη των top-k ερωτημάτων υπάρχει ένα στοιχείο για το οποίο κανένα άλλο σημείο δεν επιτυγχάνει καλύτερο σκορ ενώ για τα ερωτήματα κορυφογραμμής μπορούν να βρεθούν άνω του ενός στοιχεία για τα οποία κανένα άλλο καλύτερο στοιχείο δεν υπάρχει.

Τέλος εξετάσαμε τα ερωτήματα διαφοροποίησης αποτελέσματος (result diversification) τα οποία ενοποιούν δύο έννοιες που έρχονται σε αντίθεση. Η σχετικότητα ή ομοιότητα μίας πλειάδας ορίζεται από την απόστασή της από το ερώτημα. Από την άλλη, η διαφοροποίηση μίας πλειάδας ως προς το υπόλοιπο μέρος του αποτελέσματος ορίζεται από τον συνυπολογισμό της απόστασής της από τις εγγραφές αυτές. Η απάντηση σε ένα τέτοιο ερώτημα αποτελείται από ένα σύνολο k πλειάδων που επιτυγχάνει την καλύτερη τιμή σύμφωνα με μία αντικειμενική συνάρτηση (objective function) που συνδυάζει την σχετικότητα και τη διαφοροποίηση του αποτελέσματος. Τονίζουμε ότι εδώ κατατάσσονται σύνολα από πλειάδες κι όχι οι ίδιες οι εγγραφές κι άρα μπορεί να δειχθεί ότι το πρόβλημα είναι NP-hard.

Ειδικά τα ερωτήματα διαφοροποίησης αποτελεσμάτων αποτελούν μία σύγχρονη ανοιχτή πρόκληση με την οποία απασχολείται τα τελευταία χρόνια η επιστημονική κοινότητα. Πιο συγκεκριμένα, μελετήσαμε μία διαφορετική οπτική του προβλήματος υπό το πρίσμα κριτηρίων που αφορούν το περιεχόμενο (content-based definitions). Εναλλακτικές προσεγγίσεις του προβλήματος βασίζονται είτε στην αρχή της κάλυψης (coverage-based definitions) σύμφωνα με την οποία κάθε καινούρια πλειάδα που προστίθεται στο αποτέλεσμα καλύπτει και μία διαφορετική κατηγορία κάποιας ταξονομίας, ή της

νεωτερικότητας (novelty-based definitions), σύμφωνα με την οποία κάθε πλειάδα θα πρέπει να προσθέτει καινούρια πληροφορία στο αποτέλεσμα. Η προσέγγισή μας υιοθετεί μία Maximal Marginal Relevance (MMR) στρατηγική κατάταξης των αποθηκευμένων αντικειμένων και στηρίζεται σε ευρετικές (heuristics) τεχνικές αναζήτησης.

3.1.3 Ερωτήματα ταξινόμησης-σύζευξης

Τα ερωτήματα τύπου rank-join παίζουν σημαντικό ρόλο στις σύγχρονες διεργασίες ανάλυσης δεδομένων. Ωστόσο, παρά τη σημαντικότητά τους και σε αντίθεση με κεντρικοποιημένα περιβάλλοντα, τα ερωτήματα αυτά δεν έχουν τύχει προσοχής σε κατανεμημένα περιβάλλοντα όπως οι NoSQL βάσεων στο υπολογιστικό νέφος. Στα πλαίσια της ΥΔ 4.1 μελετήσαμε διεξοδικά ένα σύνολο κατανεμημένων big-data βάσεων και εξετάσαμε την απόδοσή τους για την αποτίμηση rank-join ερωτημάτων. Στην εργασία [NtPT414] περιγράφουμε λύσεις βάσης που χρησιμοποιούν γλώσσες τύπου SQL (όπως οι Hive και Pig) και βασίζονται σε διεργασίες MapReduce. Κατόπιν, προσφέρουμε λύσεις που βασίζονται σε εξειδικευμένες δομές δεικτοδότησης, οι οποίες μπορούν με τη σειρά τους να προσπελαστούν είτε μέσω διεργασιών MapReduce, είτε με στρατηγικές βασισμένες σε συντονιστές. Η πρώτη λύση δεικτοδότησης βασίζεται σε ανεστραμμένους δείκτες που προσπελαύνονται με MapReduce. Η δεύτερη λύση προσαρμόζει έναν δημοφιλή κεντρικοποιημένο αλγόριθμο επεξεργασίας ερωτημάτων rank-join. Επιπλέον συνεισφέρουμε μία καινοτόμα στατιστική δομή που συνδυάζει ιστογράμματα και φίλτρα Bloom, και η οποία αποτελεί τη βάση της τρίτης μας λύσης. Περιγράφουμε (α) αλγορίθμους MapReduce για την δημιουργία των δεικτών και δομών αυτών, (β) αλγορίθμους για την online ενημέρωσή τους, και (γ) αλγορίθμους επεξεργασίας ερωτημάτων που χρησιμοποιούν τα παραπάνω.

Υλοποιήσαμε όλες τις λύσεις πάνω από Hadoop (HDFS) και HBase και τις δοκιμάσαμε με δεδομένα από τη δέσμη TPC-H σε διάφορες κλίμακες και με διαφορετικά ερωτήματα σε πίνακες διαφόρων μεγεθών και με διαφορετικές κατανομές βαθμών-χαρακτηριστικών. Μεταφέραμε την υλοποίησή μας στο Amazon EC2 και σε συστάδες υπολογιστών διαφόρων μεγεθών στο εργαστήριό

μας. Παρέχουμε αποτελέσματα απόδοσης για τρεις μετρικές: χρόνο επεξεργασίας ερωτημάτων, κατανάλωση εύρους ζώνης δικτύου και "χρηματικό" κόστος για την εκτέλεση των ερωτημάτων.

Η συγκεκριμένη εργασία αφορά σε θέματα κομβικής σημασίας για την κατανεμημένη επεξεργασία κατανεμημένων δεδομένων. Ειδικότερα, προσαρμόζει τεχνολογίες και συστήματα αιχμής, όπως MapReduce και NoSQL συστήματα δεδομένων και αναπτύσσει δομές και αλγόριθμους για την εξυπηρέτηση πολύ σημαντικών σύνθετων ερωτημάτων, όπως αυτά που απαιτούν σύζευξη και διάταξη (rank-joins) που είναι στην καρδιά των δραστηριοτήτων της δράσης 4 του προγράμματος ΕΙΚΟΣ. Η εργασία είναι η πρώτη μέθοδος που δείχνει πως αυτά τα ερωτήματα μπορούν να απαντηθούν στις νέες τεχνολογίες συστημάτων Big Data. Ως τέτοια, παίζει σημαντικό ρόλο στα πλαίσια του προγράμματος ΕΙΚΟΣ.

3.2 Ανθρωποκεντρικές τεχνικές δεικτοδότησης

Στις μέρες μας οι χρήστες εκτός από "καταναλωτές" πληροφορίας σε ένα οικοσύστημα είναι και "παραγωγοί" και διαχειριστές της. Μια συνήθης πρακτική είναι η σήμανση του περιεχομένου που διαμοιράζονται με ετικέτες (tags) και η χρήση των ετικετών αυτών σε διαδικασίες αναζήτησης ή εύρεσης περιεχομένου με παρόμοια χαρακτηριστικά. Ένα από τα εργαλεία που ευρέως χρησιμοποιούνται στις διαδικασίες αυτές είναι οι ταξονομίες (taxonomies). Οι ταξονομίες είναι δενδρικές δομές που συνίστανται από κόμβους, καθένας από τους οποίους αντιπροσωπεύει μια κατηγορία-έννοια και συνδέεται με τα παιδιά και τον γονέα του με σχέσεις "IS-A". Δημιουργούνται κατά βάση χειρωνακτικά, από ειδικούς, ενώ η ανανέωση και επέκτασή τους είναι αρκετά χρονοβόρες ενέργειες.

Στα πλαίσια της ΥΔ 4.1. μελετήσαμε την αυτόματη εξαγωγή ταξονομίας από είσοδο που προέρχεται από μια κοινότητα χρηστών. Θεωρούμε πως οι χρήστες μας είναι ικανοί να παρέχουν σχέσεις ετικετών που περιγράφουν καταστάσεις υπερκατηγορίας-υποκατηγορίας μεταξύ θεματικών κόμβων και προσπαθούμε να τις συγκεράσουμε ώστε να προκύψει μια ταξονομία. Η τελική ταξονομία είναι

συμβατή με τα ‘κβάντα’ πληροφορίας που έχουμε στη διάθεσή μας, επιλύει αντικρουόμενες απόψεις χρηστών όσον αφορά την τελική της δομή και αποτυπώνει την ικανότητα της κοινότητας στη διακριτοποίηση εννοιών.

Στην εργασία [KaTr12] προτείνουμε έναν αλγόριθμο κατασκευής μιας ταξονομίας και αξιολογούμε την απόδοσή του χρησιμοποιώντας τόσο συνθετικά, όσο και πραγματικά δεδομένα. Γίνεται επίσης προσαρμογή και μελέτη του συστήματος σε crowdsourcing περιβάλλοντα, δηλαδή περιβάλλοντα όπου ένας μεγάλος αριθμός από χρήστες χρησιμοποιείται για την περάτωση μικρών εργασιών που χαρακτηριστικό τους είναι η αδυναμία εκτέλεσής τους από υπολογιστικά συστήματα. Στα πλαίσια αυτά, η παρούσα εργασία προτείνει ένα περιβάλλον για τη συλλογή και επεξεργασία πληροφορίας από τους χρήστες. Ορίζουμε τη διεπαφή που οι τελευταίοι έχουν με το σύστημα και προτείνουμε έναν αλγόριθμό για τη δημιουργία μιας δομής δεικτοδότησης. Η δομή αυτή είναι αποτέλεσμα συλλογικής εργασίας, στα πλαίσια της κοινότητας των χρηστών και αποτυπώνει τη γενική αντίληψη που έχει η κοινότητα ως προς το φυσικό περιβάλλον γύρω της.

3.3 Αντιμετώπιση αγνοούντων τιμών

Η επίλυση του προβλήματος των αγνοούντων τιμών με μικρά στατιστικά σφάλματα σε περιβάλλοντα δεδομένων μεγάλης κλίμακας (big data) αποτελεί μια μεγάλη πρόκληση. Όσο το μέγεθος των δεδομένων και η κοινότητα χρηστών μεγαλώνουν το πρόβλημα δυσκολεύει έτι περαιτέρω. Ας υποθέσουμε πως είναι δυνατόν να έχουμε μια μηχανή (“Γκοντζίλα”) που δύναται να αποθηκεύσει τα τεράστια σε όγκο δεδομένα και να υποστηρίξει μια μεγάλη κοινότητα χρηστών που υποβάλλουν αιτήσεις για την αντικατάσταση τιμών που αγνοούνται. Είναι δυνατόν να αντικαταστήσουμε τον “Γκοντζίλα” με ένα μεγάλο αριθμό από μηχανές-βοηθούς (cohorts) έτσι ώστε οι αντικαταστάσεις τιμών να γίνονται πολύ γρηγορότερα, εμπλέκοντας cohorts παράλληλα, ο καθένας εκ των οποίων προσπελαύνει πολύ μικρότερο όγκο των αρχικών δεδομένων; Αν ναι, θα ήταν προτιμότερο για λόγους καλύτερης απόδοσης, να προσπελαύνονται μόνο ένα μικρό ποσοστό των cohorts. Σε αυτήν την περίπτωση, μπορούμε να αποφασίσουμε γρήγορα ποιό είναι το καλύτερο υποσύνολο cohorts να

προσπελαστούν για κάθε αίτηση; Και να εξασφαλίσουμε καλύτερο στατιστικό λάθος από ότι ο “Γκοντζίλας”;

Στην εργασία [AnTr14] απαντούμε σε όλες τις παραπάνω θεμελιώδεις ερωτήσεις, καθιστώντας τη συγκεκριμένη δουλεία την πρώτη που μελετά συστηματικά το πρόβλημα σε κατανεμημένα κλιμακώσιμα συστήματα (scale-out). Ειδικότερα, μελετήσαμε το πρόβλημα της αντικατάστασης αγνοούμενων τιμών σε συστήματα δεδομένων μεγάλης κλίμακας και αναπτύξαμε στατιστικές δομές και αλγόριθμους για την εξυπηρέτηση αιτήσεων αντικατάστασης τιμών που είναι σημαντικά για πολλές εφαρμογές στην καρδιά των δραστηριοτήτων της Δράσης 4 του προγράμματος ΕΙΚΟΣ. Η εργασία είναι η πρώτη μέθοδος που δείχνει πως αυτά τα ερωτήματα μπορούν να εξυπηρετηθούν σε μεγάλα συστήματα scale-out στις νέες τεχνολογίες συστημάτων Big Data.

4 Ανακεφαλαίωση

Το παρόν παραδοτέο Π4.1 παρουσιάζει τα αποτελέσματα της υποδράσης ΥΔ 4.1 του έργου ΕΙΚΟΣ. Ο στόχος της υποδράσης ΥΔ 4.1 ήταν η σχεδίαση αποδοτικών τεχνικών εξατομικευμένης επεξεργασίας σύνθετων επερωτήσεων σε μεγάλης κλίμακας πληροφοριακά συστήματα. Στα πλαίσια της υποδράσης καταλήξαμε σε τρεις κεντρικούς άξονες γύρω από τους οποίους εστιάσαμε την έρευνα μας.

1. Για την αντιμετώπιση της κλιμάκωσης της ποσότητας των δεδομένων και την αποφυγή βομβαρδισμού του χρήστη με μη αξιοποιήσιμα αποτελέσματα στις αναζητήσεις του, μελετήσαμε τεχνικές που παρουσιάζουν ταξινομημένα και εξατομικευμένα αποτελέσματα σε σύνθετες επερωτήσεις. Στα πλαίσια των εργασιών της ΥΔ 4.1 εξετάσαμε τους ποιο δημοφιλής τύπους ερωτημάτων (top-k, skylines, reverse skylines, rank-joins) και αναπτύξαμε μη τετριμμένες τεχνικές συνδυασμού τους για τη προσωποποιημένη πρόσβαση σε πληροφοριακά συστήματα μεγάλης κλίμακας.

2. Για την αποδοτικότερη δεικτοδότηση των δεδομένων σε εφαρμογές μεγάλης κλίμακας αναπτύξαμε καινοτόμες ανθρωποκεντρικές τεχνικές αξιοποιώντας

ετικέτες (tags) επισήμανσης και μελετήσαμε τρόπους δημιουργίας ταξονομιών μέσω τεχνικών crowdsourcing.

3. Μελετήσαμε τεχνικές οι οποίες επιτρέπουν την αντιμετώπιση ελλιπής ή ασαφούς πληροφορίας σε περιβάλλοντα μεγάλων δεδομένων. Συγκεκριμένα αναπτύξαμε κατανεμημένες τεχνικές αντικατάστασης αγνοούμενων τιμών οι οποίες επιτρέπουν την κλιμάκωση των συστημάτων.

Στα πλαίσια των ερευνητικών δραστηριοτήτων της ΥΔ 4.1 προέκυψαν 5 συνολικά δημοσιεύσεις (βλ. παρακάτω πίνακα). Σε αυτές παρουσιάζονται αναλυτικά οι αλγόριθμοι και τεχνικές και μελετάται πειραματικά η απόδοση τους χρησιμοποιώντας αναδυόμενες τεχνολογίες συστημάτων μεγάλων δεδομένων.

Δημοσιεύσεις

[ArDV12] Anastasios Arvanitis, Antonios Deligiannakis, Yannis Vassiliou. Efficient influence-based processing of market research queries. CIKM 2012: 1193-1202

[TsSS14] George Tsatsanifos, Dimitris Sacharidis, Timos Sellis. RIPPLE: A Scalable Framework for Distributed Processing of Rank Queries. EDBT 2014: 259-270

[KaTr12] Dimitris Karampinas, Peter Triantafillou. Crowdsourcing Taxonomies. ESWC 2012: 545-559

[NtPT14] Nikos Ntarmos, Ioannis Patlakas, Peter Triantafillou: Rank Join Queries in NoSQL Databases. HDMS 2014

[AnTr14] Christos Anagnostopoulos, Peter Triantafillou. Scaling out big data missing value imputations: pythia vs. godzilla. KDD 2014: 651-660

Παράρτημα